A method to computationally screen for tunable properties of crystalline alloys

Rachel Woods-Robinson*[1,2], Matthew K. Horton*[2,3], and Kristin A. Persson[3,4]
[1]*Applied Science and Technology Graduate Group, University of California at Berkeley, Berkeley, CA, 94720 USA,*
[2]*Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720 USA,* [3]*Department of Materials Science and Engineering, University of California at Berkeley, Berkeley, CA, 94720 USA,*
[4]*Molecular Foundry Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720 USA,* *These authors contributed equally to this research.*
(Dated: June 23, 2022)

Conventionally, high-throughput computational materials searches start from an input set of bulk compounds extracted from material databases, and this set is screened for candidate materials for specific applications. In contrast, many functional materials, and especially semiconductors, are heavily engineered alloys of multiple compounds rather than a single bulk compound. To improve our ability to design functional materials, in this work we propose a framework to automatically construct possible "alloy pairs" and "alloy systems" and detect "alloy members" from a set of existing, experimental or calculated ordered compounds, without requiring any additional metadata beyond their crystal structure. As a demonstration, we apply this framework to all inorganic materials in the Materials Project database to create a new database of over 600,000 unique "alloy pair" entries which can then be used in materials discovery studies to search for materials with tunable properties. This new database has been incorporated into the Materials Project website and linked with corresponding material identifiers for any user to query and explore. Using an example of screening for p-type transparent conducting materials, we demonstrate how using this methodology reveals candidate material systems that might otherwise have been excluded by a traditional screening. This work lays a foundation from which materials databases can go beyond stoichiometric compounds, and approach a more realistic description of compositionally tunable materials.

arXiv:2206.10715v1 [cond-mat.mtrl-sci] 21 Jun 2022

INTRODUCTION

The power of functional semiconductor materials lies in the tunability of their properties. Since the dawn of the Semiconductor Age, traditional semiconductors—elemental (e.g. Si), IV-IVs (SiC), III-Vs (GaN, GaAs, InGaN), II-IVs (CdTe), etc.—have been manipulated in the laboratory through doping, alloying, processing, and other techniques to yield desired properties. Tunable semiconductor alloy materials enable a variety of energy and optoelectronic applications that govern our modern world, from light-emitting diode (LED) materials e.g. $In_{1-x}Ga_xN$[1] to infrared detectors e.g. $Pb_{1-x}Sn_xTe$ and $Hg_xCd_{1-x}Te$[2] to piezoelectrics e.g. $PbZr_xTi_{1-x}O_3$,[3] and are critical for the transformation to renewable energy in solar cell materials e.g. $CuIn_xGa_{1-x}Se_2$ (CIGS)[4] and $CdSe_xTe_{1-x}$ (CdTe). The properties of each of these materials reach far beyond those of their endpoint compositions, e.g., the band gap of InGaN is tunable across a wide range, from ∼0.7 eV (InN) to 3.4 eV (GaN). Naturally occurring semiconductor minerals are also stable in alloy forms, e.g. olivine $(Mg_xFe_{1-x})_2SiO_4$, plagioclase $Na_xCa_{1-x}(Al_ySi_{1-y})_4O_8$, and cobaltite $Co_xFe_{1-x}AsS$, indicating a strong tendency towards off-stoichiometric stability.[5]

Meanwhile, in the past decade computational materials discovery has been advancing novel materials design in a wide range of applications, from thermoelectrics,[6] to Li-ion battery cathodes,[7] to transparent conductors.[8] In most of these cases, materials discovery has been targeted towards stoichiometric "bulk" compounds (also called "parent" compounds or "endpoint" compounds in the context of alloys). A candidate compound emerges successfully from a screening if it satisfies a set of property values within a specific cutoff. This methodology has served as a useful starting point, but a grand challenge in the field is determining how to expand this success beyond compounds into off-stoichiometric space to search for *ranges of tunability* within materials in a high-throughput context. Indeed, a material may be excluded by its endpoint properties without taking into account how its properties can be tuned by doping or alloying. For example, the n-type transparent conductor Sn-doped In_2O_3 is an excellent example of a material where the computed properties of the endpoint compound (In_2O_3) are not representative of the high experimental performance achieved by introducing tunability.[9]. It is recognized that considering all possible off-stoichiometry (defects, dopants, impurity phases and alloying) — intentional as well as unintentional — in the design of novel materials incurs a vast increase in complexity of search space as compared to on-stoichiometric compound space. Therefore, part of the challenge is a data problem: how do we manage the additional complexity induced by including off-stoichiometry?

There have been many extensive and notable previous efforts to designing alloys using high-throughput computation. These include but are not limited to the design of high-entropy alloys[10, 11], high-entropy oxides[12], Heusler compounds[13] and magnetic Heuslers [14], as well as alloy design for specific applications including magnetocalorics[15] and thermoelectrics[16]. Such design studies often bootstrap alloy searches from existing computational databases, such as the Materials Project, AFLOW[17] and OQMD[18]. Previous efforts have also used novel approaches[19, 20] includ-

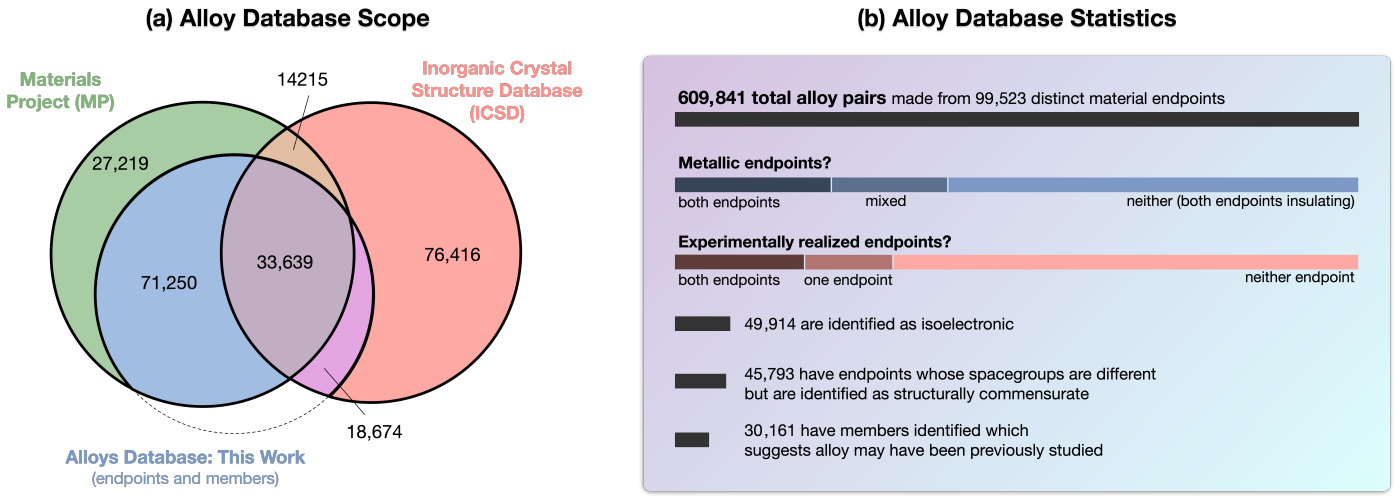

Figure 1: (a) A Venn diagram showing materials that are (i) in the Materials Project (MP), (ii) that are associated with an alloy in the alloy database presented in this work (whether an alloy endpoint or an alloy member), and (iii) are in the Inorganic Crystal Structure Database (ICSD). Note that each entry in the Materials Project represents a distinct polymorph, whereas duplicates are present in the ICSD, and so the ICSD is likely overcounted. The Venn diagram gives an overview of how these three databases relate to each other. (b) A summary of statistics within the alloy database presented in this work.

ing machine learning[21] and DFT-supported CALPHAD methodologies[22]. The importance of considering alloys in high-throughput computation is therefore well-known.[23] However, what many of these prior examples have in common is that they are often focused on the generation of new alloy materials within a limited regime of phase space; this is often from the enumeration of possibilities from a single crystal structure prototype, or is limited to binary alloys or a restricted chemical space. In contrast, our current work differs in that it offers a general approach for classifying and searching existing databases, such as those that typically arise from high-throughput computational studies, and therefore yields possibilities for more effective materials discovery screenings.

To clarify the scope of this work, we will recap what we mean by "alloy" in this context. The Hume-Rothery rules,[24] traditionally applied to metals, provide a guideline for considering whether two materials (A and B) may form a substitutional solid solution with each other (A_xB_{1-x}), whereby one atom is replaced by another but the host lattice remains largely unchanged, except for small local distortions. These rules require that (1) that the crystal structures of solute and solvent must be similar (that is, commensurate with each other), (2) that the atomic radius of solute and solvent atoms must differ by no more than 15%, (3) that solvent and solute have the same valency for complete solubility, and (4) the solute and solvent should have similar electronegativity. These rules are good guidelines, although the cutoffs ("15%", "similar") are open to debate. The methodology presented in this work therefore is focused primarily on rule (1) to generate the database using existing algorithms for assessing crystal structure similarity, with sufficient metadata then retained to assess rule (3) by querying the database. Rules (2) and (4) are easily applied by the person retrieving alloys from the database subject to their own materials design requirements; for example, by accessing the database of ionic radii within `pymatgen` to further filter down the list of possible alloys to consider. We emphasize that the alloy database obtained in this work is only a database of *possible* alloys with respect to these rules, and does not guarantee that these alloys do indeed exist.

Using this database, we create methodology to aid in the analysis of alloying opportunities, enabling computational screening for tunable properties in inorganic alloys when starting from a database of crystallographic structures and associated properties. First, we map tunable material space and search for substitutional alloy compositions and properties within a given set of possible endpoints. Second, we apply this framework to the entire Materials Project (MP) database[25] for commensurate[26] (structure-matching within a certain tolerance) structures to enable analysis resulting in over 600,000 alloy tielines, encompassing 270,545 chemical systems and 215 space groups. Third, we provide a series of new techniques to conceptualize and explore this large alloy space, including defining an "alloy system" comprised of alloy pairs, thermodynamic stability estimates of alloys by alloy content using a "half-space hull" approach, and an example of a high-throughput screening for alloy pairs. And lastly, we outline the limitations of this framework, and suggest next steps for tunable material screenings.

We focus on semiconductors in this paper, but the general methodology could be applied to any alloy systems where there is a reasonable expectation of structural stability and approximately linearly-dependent properties with composition. The alloy framework developed in this work is available in the open-source `pymatgen-analysis-alloys` repository and the analyses and associated enabling functionalities have been incorporated into the Materials Project website under a Creative Commons license, with an API to enable other researchers

to explore the data and download the results. We note that this API will be publicly available subsequent to the publication of this pre-print after peer-review.

RESULTS AND DEMONSTRATIONS

Creating an alloy database of "alloy pairs"

In brief, our methods combine sets of structurally commensurate endpoint compounds into an "alloy pair" database record, which represent two compositions with the possibility of forming a solid solution with one another (see Methodology and SI for details). For example, endpoint compounds wurtzite GaN and wurtzite InN form an alloy pair $Al_xGa_{1-x}N$. Applying the methodology described here to the Materials Project database produces 609,841 alloy endpoint pairs. Of these candidate alloy systems, 30,161 pairs (4.9%) and 2,120 (17.9%) systems are found to contain members of intermediate, non-stoichiometric "member" compositions, suggesting that these may have been previously explored either experimentally or computationally. **Figure 1** depicts a summary of the dataset, as a subset of both the Materials Project (MP)[25] and Inorganic Crystal Structure Database (ICSD),[27] and is broken down by categories including whether the alloy is metal–metal, metal–semiconductor, or semiconductor–semiconductor, and whether the alloy endpoints have been previously synthesized experimentally.

In Figure 1, we also highlight that we have determined 45,793 alloy pairs whose endpoint compounds are not detected to have the same space group. This can either be because the detected space group, being subject to numerical tolerances, is incorrect, or it can be a sign of a phase transition. An instance of the latter case might be, for example, one endpoint of an alloy pair might have a small polar distortion, while the other endpoint might be a nonpolar material; here, the space groups of the endpoints do not match, but the materials might still be "commensurate" and able to alloy. This demonstrates the importance of carefully selecting the method for which two materials are considered to be structurally commensurate, and so might form a substitutional alloy. In the context of a materials screening, including alloys drastically expands the accessible and searchable parameter space. When properties of an alloy pair are considered, we take properties of the end-points when known and assume Vegard's law with no bowing for lattice constant, E_G and $1/m_h^*$.[28] We note that excluding bowing is a crude approximation for band gap, but bowing is not as significant for effective mass (see SI).[29]

Exploration of alloy systems

By combining alloy pairs that are all commensurate with one another, "alloy systems" can be generated (see Methodology) in which each alloy system spans a region of accessible phase space. Applying this methodology to the Materials Project creates a total of 11,876 possible alloy systems. One application of the alloy system framework is the construction of semiconductor bowing plots, which are useful for visualizing lattice matching and band gap tuning in semiconductor alloys, and are typically constructed manually via a literature review. A typical example might be a plot showing wurtzite III-V alloys system (GaN, InN, etc.), but can be generalized for any alloy system. Here, we take an example of two systems which have been studied experimentally but not as extensively as the III-V system: zincblende II-*Ch* and chalcopyrite I-III-Ch_2 chalcogenide materials [30] Compounds are grouped by commensurate structure, each marker corresponding to an experimentally observed endpoint compound, and each tieline corresponds to an experimentally observed alloy (e.g. $Zn_xMg_{1-x}S$). For most of the compounds plotted are in their most stable polymorph, however we note some exceptions (e.g., MnN, MnSe, and CdSe have a more stable polymorph than zincblende, but zincblende is plotted here for clarity and completeness).

Using the alloys systems framework, we generate corresponding alloy systems for zincblende and chalcopyrite chalcogenide semiconductors. These systems are plotted in the two panels of **Figure 2**(b) as a function of lattice parameter a and band gap $E_G^{\mathrm{PBE,corr}}$, where each commensurate system is merged into a shape to represent their range. Alloy systems are generated as a function of a single compound — in this case zincblende ZnS and chalcopyrite $CuAlS_2$ — and then outputs are filtered to include only chalcogenide compounds with commensurate oxidation states. Since semi-local DFT underestimates the band gap, we plot PBE gaps with an applied approximate empirical correction factor from the literature, as $E_G^{\mathrm{PBE,corr}}$.[31] However, for better accuracy, we recommend performing additional hybrid functional calculations to complement the initial screening and provide a better estimate the gap in the alloy database or, in the future, using more accurate calculations to construct the database. We emphasize that the purpose of this work is not to demonstrate accurate band gap prediction since more accurate methods are well-known, but to demonstrate the machinery of constructing alloy pairs and connecting these into alloy systems for the purposes of a materials discovery screening.

We observe in Figure 2 that the shapes and features of computationally-generated alloys systems in (b) qualitatively match the experimental diagrams in (a), subject to uncertainities in band gaps as explained above. Additionally, more information is captured in (b), in particular the members (MP and ICSD) of many of the tielines are denoted which indicate which alloys have seen previous study. Including additional hypothetical alloy pairs here increases range of search space, by nearly 50% percent for II-*Ch* and by over 50% for I-III-Ch_2, and new alloy pair tielines are marked with dotted lines such as $Ca_xCd_{1-x}Se$ and $AgAl(Se_xTe_{1-x})_2$. The computed alloy system plots can also inspire new materials design searches over a variety of multinary alloys. For example, in a search an am-

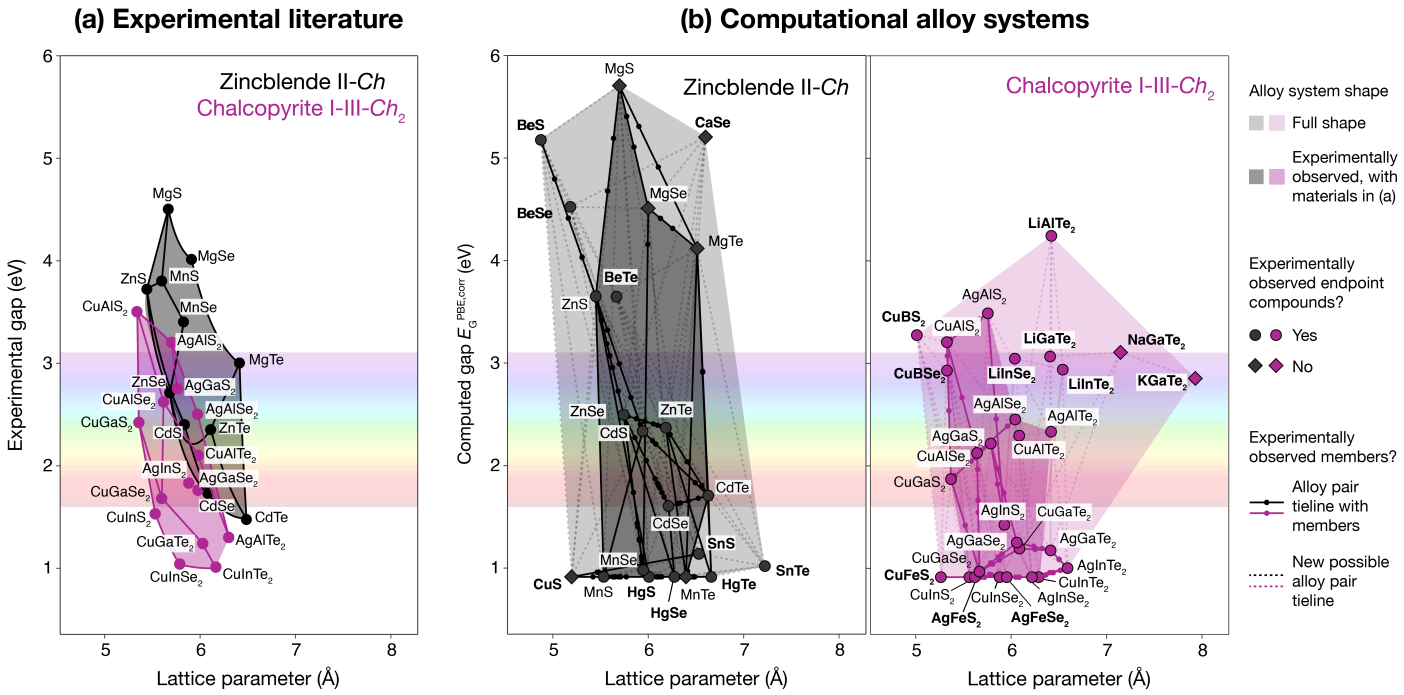

Figure 2: A comparison of a lattice parameter versus band gap plot for wide band gap zincblende and chalcopyrite chalcogenide materials (a) a manually-constructed plot from the experimental literature,[30] including experimental bowing, and (b) generated computationally from an "alloy system" in our database. Note that in (b) each alloy system is filtered to include only chalcogenide (S, Se, Te) compounds with commensurate oxidation states, that band gaps $E_{\mathrm{G}}^{\mathrm{PBE,corr}}$ are computed PBE gaps with an approximate empirical correction factor,[31] and that the lattice parameter is computed from the conventional unit cell.

ber LED material (∼580-590 nm, i.e. 2.10–2.14 eV) with a lattice parameter matched to zincblende GaAs (5.6531 Å), one may examine the region around $Zn_xMn_{1-x}S$ or $CuAl_xGa_{1-x}Se_2$ alloy pairs.

In principle, for a given alloy system, an alloy's lattice parameter (or volume cube root, if comparing non-cubic systems) and band gap can be tuned within the bounds of the shape bounded by the alloy end-members in the plot by varying alloy composition. Here we show a plot for a simple comparison with conventional semiconductor a vs. E_{G} plots, but alloy system plots can be created for any set of properties and can in principle be expanded into higher dimensions. Some degree of bowing is likely in these systems, as shown in Figure 2(a). Additionally, discontinuities in Vegard's law can arise when gaps transition from direct to indirect nature across alloy space. However, this analysis is helpful as an initial guide and provides a systematic method for generation of such figures which are already commonplace in the literature.

Estimating alloy stability

So far we have introduced a set of alloys that are purely hypothetical, yet have not discussed thermodynamic stability and synthesizability. Computationally determining synthesizability of a given alloy is non-trivial, and requires constructing a temperature-dependent phase diagram, computing the effects of various entropy terms, and potentially considering the effects of nucleation and kinetics. A simple (yet imperfect, due to the typical T = 0K approximation of DFT) metric to assess stability in solid state compound materials is the "energy above the convex hull" (E_{hull}), where an E_{hull} of 0 eV/atom defines a thermodynamic ground state, and E_{hull} is reported for compounds in the MP database as derived from phase diagrams constructed from DFT calculations. For alloys, even determining this simple E_{hull} metric is nontrivial — it requires computing a variety of orderings, as well as fully exploring possible competing polymorph phases and all of their possible orderings (see Discussion).

Here, as a method of approximating whether a given alloy may be stable or synthesizable, and a first step before performing additional in-depth calculations, we have calculated a hull across alloy content using the half-space intersection of the lines representing the linear interpolation of formation enthalpies between the two alloy endpoints — hereafter called a "half-space hull" —t o identify ranges of alloy content x at which different polymorphs might be stable. This is defined by a "formula alloy pair" and "alloy segments" (see Methodology).

Figure 3(a) depicts a set of "formula alloy pairs" derived from the alloy database, made up of alloy pairs that all have the same composition. For example, a $Sb_xBi_{1-x}OF$ alloy is shown in the third panel of (a) and is magnified in Figure 3(b), with BiOF endpoint compounds on the left side (x=0) and SbOF endpoint compounds on the right side (x=1), and with the y-axis representing E_{hull}. Only compounds that are present in the MP database are included here. For each case where a BiOF compound is

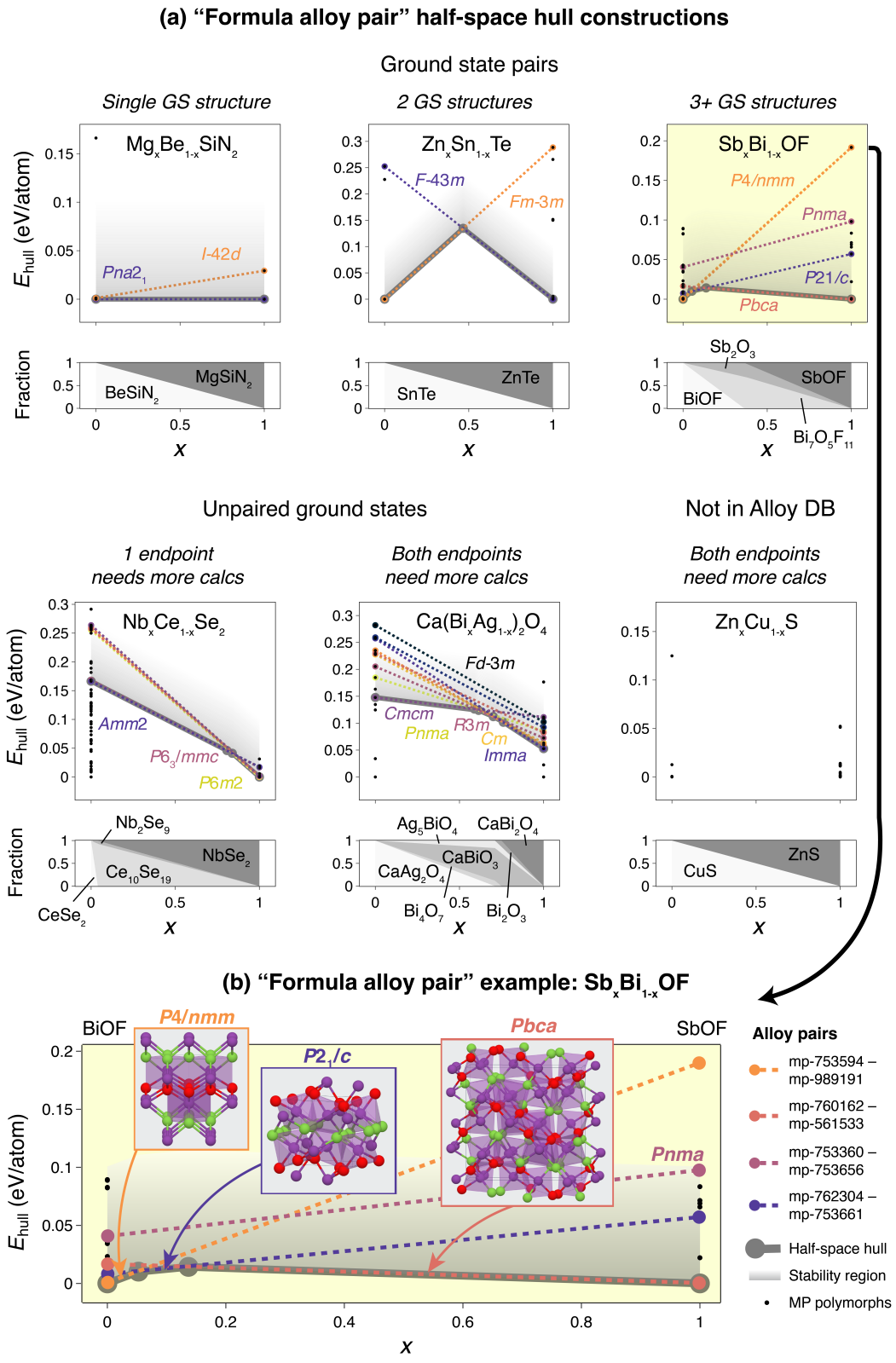

Figure 3: (a) A set of half-space hull intersections, a simple interpolation based on endpoint formation energies to find cross-overs, for six representative formula alloy pairs. It is not expected these cross-over points will be exact but might provide an estimate. For each alloy system, this then gives a range of allowed compositions and phases. Below each half-space hull construction, a decomposition diagram is plotted as a function of x. (b) A representative formula alloy pair of $Sb_xBi_{1-x}OF$, with three unique phases lying on the half-space hull. Dotted lines depict where a competing polymorph comes stable. Decomposition plot on the right shows thermodynamic decomposition products from a ternary phase diagram, as a function of x.

structurally commensurate with a SbOF compound, an `AlloyPair` is formed and a colored dotted line is drawn in Figure 3(b). For example, $P4/nmm$ BiOF (mp-753594, on the hull) is connected with $P4/nmm$ SbOF (mp-989191, with E_{hull}=0.192 eV) by a blue dotted line, while $Pcba$ BiOF (mp-760162, with E_{hull}=0.017 eV) is connected with $P4/nmm$ SbOF (mp-561533, on the hull) by a green dotted line. In this formula alloy pair, $P2_1/c$ (purple) and $Pnma$ (red) pairs are also drawn.

The "half-space hull" is drawn as a continuous grey line in Figure 3(b). The changes of slope along this line represent "alloy segments" (see Methodology), which represent phase changes as x is increased. Thus, in this example the half-space hull defines segments of $Sb_xBi_{1-x}OF$ where $P4/nmm$ is the lowest energy phase ($0 \lesssim x \lesssim 0.05$), where $P2_1/c$ is the lowest energy phase ($0.05 \lesssim x \lesssim 0.15$), and where $Pbca$ is the lowest energy phase ($0.15 \lesssim x \lesssim 1$). Since a phase does not have to lie on the hull to be synthesizable, we draw a region above the half-space hull (the "stability region" in a shaded grey gradient) at which the "energy above the half-space hull" is less than 0.1 eV/atom. It is typical in materials screenings to define an arbitrary cutoff such as this, below which materials are more likely to be synthesizable or metastable. According to this cutoff, it may be possible to synthesize alloys that lie within the grey region, rather than only the alloys that lie directly upon the half-space hull. For example, it may be possible to synthesize $Pbca$ $Sb_xBi_{1-x}OF$ at small values of x, however it is far less likely to be able to synthesize $P4/nmm$ $Sb_xBi_{1-x}OF$ alloys at high values of x since the tieline for this alloy pair lies well outside of the stability region. We note that there are other endpoint compounds that do not have commensurate pairs (black circular markers), and for this method to be technically complete the formation energies of the commensurate structure pairs for these polymorphs would have to be computed.

Figure 3(a) depicts other possible scenarios of formula alloy pairs within the alloys database. Cases where both endpoints have paired ground states (three examples on the left) are most likely to provide useful information using the half-space hull method. For example, in $Mg_xBe_{1-x}SiN_2$, both ground states are $Pna2_1$ and thus it is likely that a solid solution can be synthesized across all values of x with this structure retained. In $Zn_xSn_{1-x}Te$, both endpoints have commensurate ground states and no other known polymorphs with E_{hull} <0.1 eV/atom. Thus a phase change from $Fm\bar{3}m$ to $F\bar{4}3m$ is expected at approximately x=0.5 using the half-space hull formalism. However, there are systems where one or both of the ground states do not have a commensurate pair (three examples on the right), such as $Nb_xCe_{1-x}Se_2$ and $Ca(Bi_xAg_{1-x})_2O_4$, and thus more calculations are needed in order to construct a reliable half-space hull. We note that this would add more possibly unstable or unsynthesizable endpoints.

Below each formula alloy pair in (a) is a fractional decomposition diagram. This consists of the various thermodynamic decomposition products and their fractional ratio, as a function of x, and is computed by *pymatgen*. The decomposition products gives an indicator of possible competing phases across the alloy tieline that may impede the formation of a solid solution, derived from the already phase diagram for the appropriate chemical system. In the $Mg_xBe_{1-x}SiN_2$ and $Zn_xSn_{1-x}Te$, the decomposition products consist solely of the endpoint compounds, and increases monotonically with x. However the fractional decomposition of $Sb_xBi_{1-x}OF$, enlarged and plotted on the right-hand panel of (b), is more complicated and consists of four decomposition products: endpoints BiOF and SbOF, as well as Sb_2O_3 and $Bi_7O_5F_{11}$. Thus, although the half-space hull tielines lie below 0.1 eV/atom, these $Sb_xBi_{1-x}OF$ alloys may be challenging to synthesize due to competing thermodynamic reaction products.

As a check to whether the half-space hull is appropriate as a screening tool — or in other words, whether the linearly-interpolated half-space hull estimate is consistent with the DFT computed convex hull of known alloy members — we can include members on these plots for systems in which members are present in databases and their E_{hull} values are known. For example, in **Figure 4**(a) we showcase members in the formula alloy pair construction for $Ga_xAl_{1-x}N$. It is shown that the calculated formation enthalpy of the wurtzite (space group $P6_3mc$) alloy members lie below the zincblende (space group $F\bar{4}3m$), which is consistent with with the half-space hull; here, these data points refer to the formation enthalpies as calculated with DFT using small ordered approximations from entries already existing in the Materials Project database. In Figure 4(b–g), we plot six other examples of formula alloy pair half-space hull constructions for which there are members included in the alloy pairs. For $Zr_xHf_{1-x}O_2$ (f) the alloy formation energies for space groups $P4_2/mnm$, $P4_2/nmc$, and $Fm\bar{3}m$ at x=0.5 lie nearly exactly on the dotted lines from the formula alloy pair tielines. Other systems (e.g. $Sr_xCa_{1-x}TiO_3$ and $CuLi_xNi_{1-x}O_2$) have alloys ranked in the same order as the half-space hull prediction, albeit not precisely on the predicted lines. We note that some "alloys" are extensively sampled in the Materials Project database such as $Li_2Mn_xCo_{1-x}O_3$, likely due to its interest as a battery material leading to a large amount of calculations performed on this compound with varying degrees of lithiation. All of these plots are generated using the tools provided by `pymatgen-analysis-alloys`, and can be similarly constructed for any system of interest.

Therefore, when analyzing the set of alloy pairs or alloy systems within our database, it is important to assess which half-space hull scenario a given formula alloy pair lies within and whether there exists a region of phase space where stabilization is likely. Additionally, assessing possible decomposition products informs whether to expect multiple decomposition products, which could impede formation of the alloy. Overall, the half-space hull framework of drawing lines to estimate segments of phase stability is not rigorous, since formation enthalpy does not follow Vegard's law and configurational entropy is not taken into account. Rather, this method is intended to provide an estimate of what alloys might be present and where in alloy

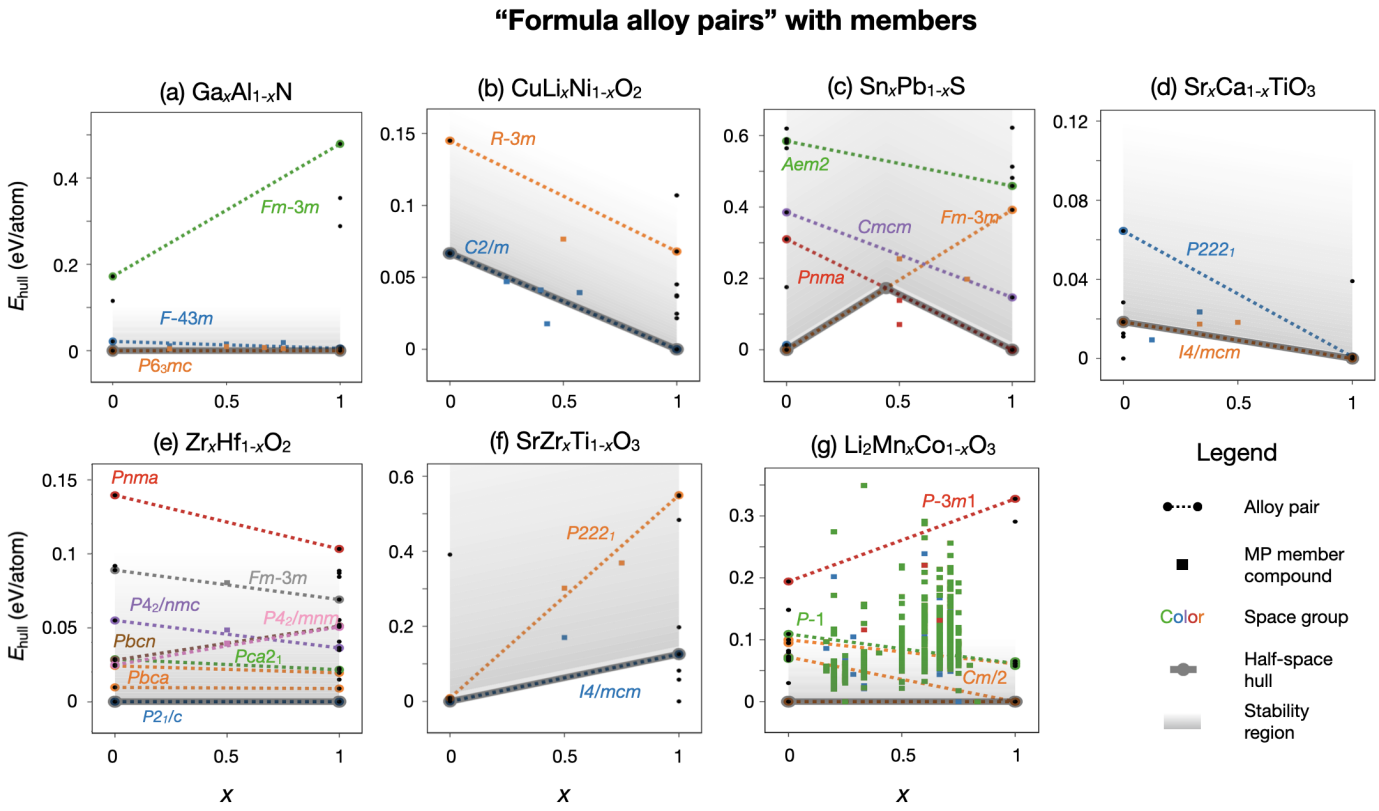

Figure 4: Examples of seven representative "formula alloy pairs" with members included. The E_{hull} of each member is sourced from the Materials Project database, and can be compared to the linearly-interpolated formation energy for each alloy pair. Alloy pairs (dashed lines) and members of alloys pairs (square markers) are colored by space group, and plots are as described in Figure 3.

space they might be, as a tool to justify or prioritize additional calculations in a high-throughput context, and is therefore an entry-point for determining which alloys may be experimentally realizable.

Example of screening alloy pairs for p-type transparent conductors

Including alloys can expand the number of material candidates generated by high-throughput screenings, and reveal candidates that otherwise would not have emerged. Here, to demonstrate this quantitatively, we screen our candidate alloy pair dataset for possible p-type transparent conductor (TC) candidates. Discovery of a high-performance p-type TC could enable breakthroughs in solar cells and transparent electronics, among other applications, but to date there are no p-type TCs that perform as well as n-type TCs.[32] A high-performing p-type TC is likely to require a low hole effective mass (m_{h}^*) to enable high hole mobility and a wide band gap (E_{G}) to enable optical transparency, among other properties.[9] So far, several data-driven explorations have been performed to search for p-type TC candidates,[8, 33, 34] but to our knowledge no screenings have been performed looking specifically for alloys or specifically for tunable materials rather than compounds.

For this analysis, we use a set of stable or metastable ($E_{\text{hull}} < 0.1$ eV) representative compounds (i.e. alloy endpoints) where m_{h}^* has been calculated (same as the set shown in Figure S1).[35] First, **Figure 5**(a) shows a set of bulk compounds from the Materials Project database, with material properties $E_{\text{G}}^{\text{PBE}}$ on the x-axis and m_{h}^* on the y-axis. The grey "p-type TC regime" depicts a range of parameter space where 1.5 eV $< E_{\text{G}}^{\text{PBE}} <$ 3 eV (and thus possible experimental gaps greater than 3 eV, since PBE systematically underestimates E_{G}) and $m_{\text{h}}^* < 1.5$, where p-type TC candidates may reside.[9] This figure plots approximately 1,000 compounds, approximately 150 of which lie within the p-type TC region and contain compounds that have emerged from previous screenings (e.g. ZrOS, $TaCu_3S_4$, and Al_2ZnTe_4). The choice of cutoff value tends to be motivated by expected values of physical parameters (e.g. absorption edge and hole mobility), but incur uncertainties in calculated value and inconsistencies between descriptor value and real physical value. Hence, the goal is to suggest a list of target candidates that may be suitable to prioritize for future computational study and experimental inquiry. Therefore, **Figure 5**(a) represents a conventional materials discovery screening.

In contrast, Figure 5(b) depicts a subset of alloy endpoint compounds (black circular markers) and corresponding alloy pair tielines (thin lines between points), and assumes Vegard's law to linearly extrapolate tielines. This analysis yields 233 alloy pairs with tielines that intersect the "p-type TC regime," and a subset of 192 alloy pairs

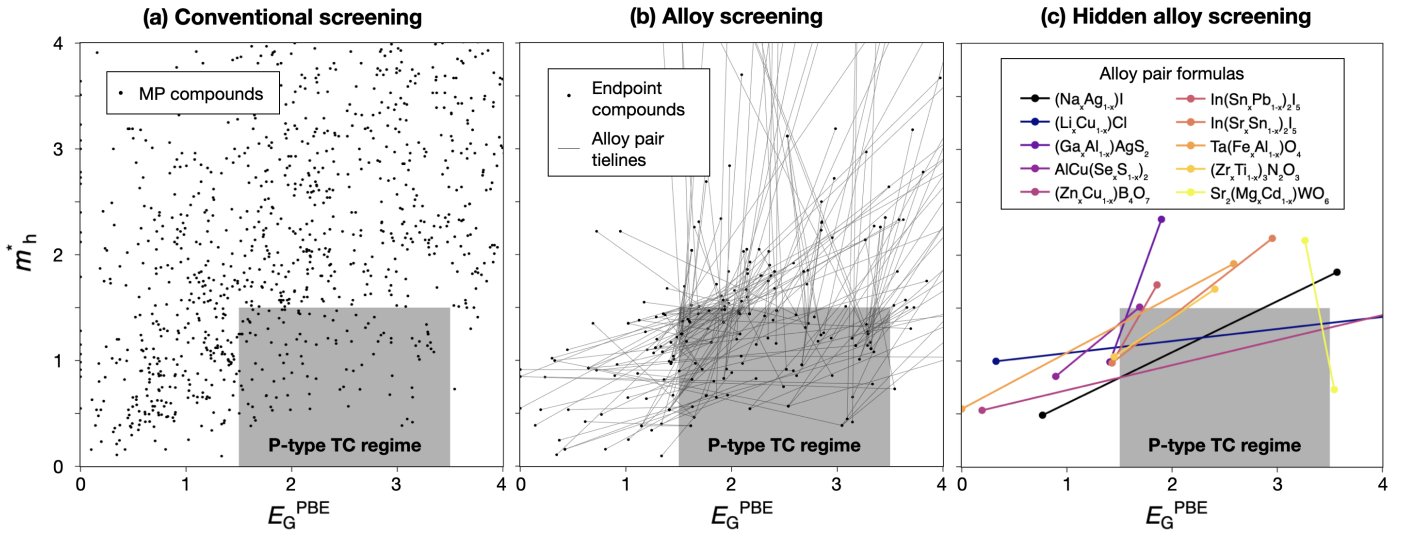

Figure 5: An example of a computational screening of an alloy search space, with the approximate computed p-type TC regime designated with a grey box. (a) All bulk compounds that intersect the approximate p-type TC regime. (b) All alloy pairs that intersect the approximate computed p-type TC regime. (c) "Hidden" alloy pairs that intersect the p-type TC regime where both endpoints lie outside of the regime. Pairs are denoted by the range of their fractional alloy compositions which lie within the regime, with details denoted in Table I.

in which one or more endpoint lies outside the regime are plotted here for readability. Thus, this plot demonstrates a set of possible, additional alloy pairs to consider as p-type TCs that previously may have been overlooked. Within the grey region, the alloy tielines indicate there may be combinations of E_G^{PBE} and m_h^* beyond those represented by the endpoint, alloy pair compounds in Figure 5(a).

In Figure 5(c), we take this a step further by highlighting a subset of ten "hidden" alloy pairs that intersect this p-type TC regime but where both of the endpoints lie outside of the regime. This analysis illustrates compounds that themselves are not p-type TC candidates but whose alloys may warrant further exploration. Table I reports all the hidden pairs from this analysis, including the ten hidden pairs from Figure 5(c). Included in this table the range of x where properties lie within the p-type TC regime ("x range"), the range of E_G^{PBE} and m_h^* achieved within this window, and E_{hull} of the endpoints (where E_{hull}^A corresponds to the first compound of a pair and E_{hull}^B to the second). It is also denoted whether a region of the x range lies on the half-space hull, and the number of decomposition products (excluding the endpoint compounds from the count). Most of the alloy pairs that emerge from this screening are quaternaries (alloys of two ternary compounds, e.g. $AlCuS_xSe_{1-x}$), with several ternaries (alloys of binary compounds, e.g. $Cu_xLi_{1-x}Cl$) and quinternaries (alloys of quaternary compounds, e.g. $Sr_2Mg_xCd_{1-x}WO_6$). To our knowledge, none of these alloy pairs have been studied previously as p-type TCs, with the exception of La_2SeO_2 and Gd_2SeO_2 which have been predicted previously using a high-throughput approach.[33] We note that this is just one example of an application where including alloying could yield new material candidates.

DISCUSSION

We have demonstrated a framework to propose new alloys and access the potential tunability of materials for high throughput screenings. In our presented database, we designate "alloy pairs" between commensurate endpoint structures; although we present 600,000 unique pairs, this database comprises a subset of possible physical alloys. Several extensions of the presented alloy database are possible, beyond constructing structure-matched pairs. For example, in many experimentally observed alloy systems, endpoints may not structure match within the tolerances we use here but are still "commensurate" with one another, i.e. they can be connected through a displacive phase transformation (e.g., orthorhombic SnS and rocksalt CaS).[26] These pairs are not included in this database, however, advances in methodologies for determining whether displacive phase transformations are possible between a given pair of materials could allow the database to be expanded in future.[36, 37] In some cases incommensurate structures, where symmetries are distinct from one another but can be connected through a reconstructive transformation, can also form heterostructural alloys which are of increased interest for materials design (e.g., rocksalt MnO and wurtzite ZnO can alloy to form $Mn_xZn_{1-x}O$).[26] Similarly, a material might be tuned by varying vacancy concentration topotactically (e.g. NiO_x). Furthermore, there are alloy pairs and alloy systems that in principle could alloy, but have no commensurate endpoint structures currently on MP (e.g. formula alloy pairs labeled "unpaired ground states" and "not in DB" in Figure 3, so in these systems more calculations would be required before the alloy could be defined. Nevertheless, in principle, the methodology presented here could be expanded upon to include and categorize all plausible commensurate and

Table I: "Hidden" alloy pairs with properties of interest to p-type TCs.

Pair IDs (A–B)	Alloy formula	Space group	x range	E_G^{PBE} range (eV)	m_h^* range	E_{hull}^A (eV/at.)	E_{hull}^B (eV/at.)	On half-space hull?†	# decomp. products‡
mp-22919–mp-23268	$(Na_xAg_{1-x})I$	$Fm\bar{3}m$	0.27–0.74	1.52–2.85	0.85–1.49	0.093	0.000	yes	0
mp-571386–mp-22905	$(Li_xCu_{1-x})Cl$	$Fm\bar{3}m$	0.20–0.52	1.51–3.46	1.13–1.36	0.178	0.020	no	1
mp-684712–mp-32891	$(Y_xGd_{1-x})_2S_3$	$I\bar{4}2d$	0.54–0.57	1.50–1.53	1.49–1.50	0.022	0.036	no	0
mp-5782–mp-556916	$(Ga_xAl_{1-x})AgS_2$	$I\bar{4}2d$	0.62–0.81	1.50–1.59	1.24–1.50	0.000	0.003	yes	0
mp-4979–mp-8016	$AlCu(Se_xS_{1-x})_2$	$I\bar{4}2d$	0.02–0.24	1.50–1.68	1.35–1.50	0.000	0.000	yes	0
mp-756317–mp-3536	$Al_2(Mg_xHg_{1-x})O_4$	$P4/mbm$	0.05–0.16	1.53–1.94	1.21–1.49	0.087	0.000	yes	1
mp-756317–mp-2908	$Al_2(Zn_xHg_{1-x})O_4$	$P4/mbm$	0.07–0.62	1.52–2.91	1.13–1.50	0.087	0.000	yes	1
mp-9081–mp-11742	$CsNd(Te_xS_{1-x})_2$	$R\bar{3}m$	0.55–0.75	1.50–1.67	1.36–1.49	0.002	0.000	yes	0
mp-555093–mp-558690	$(Zn_xCu_{1-x})B_4O_7$	$Cmcm$	0.26–0.63	1.53–3.48	0.85–1.31	0.047	0.058	yes	1
mp-13973–mp-7233	$(La_xGd_{1-x})_2SeO_2$	$P\bar{3}m1$	0.14–0.15	1.50–1.51	1.50–1.50	0.000	0.000	yes	0
mp-23520–mp-23417	$In(Sn_xPb_{1-x})_2I_5$	$I4/mcm$	0.31–0.83	1.50–1.73	1.11–1.49	0.056	0.023	yes	2
mp-23417–mp-23504	$In(Sr_xSn_{1-x})_2I_5$	$I4/mcm$	0.05–0.44	1.50–2.09	1.04–1.50	0.023	0.046	yes	2
mp-754818–mp-756933	$(Tl_xNa_{1-x})TaO_3$	$P4/mbm$	0.36–0.53	1.50–1.91	1.39–1.50	0.087	0.002	yes	0
mp-7482–mp-8402	$Rb(Mg_xHg_{1-x})F_3$	$Pm\bar{3}m$	0.14–0.16	1.52–1.64	1.41–1.47	0.000	0.002	yes	0
mp-760396–mp-761390	$Ta(Fe_xAl_{1-x})O_4$	$I4_1md$	0.31–0.42	1.51–1.79	1.35–1.50	0.056	0.019	no	0
mp-755054–mp-755998	$(Zr_xTi_{1-x})_3N_2O_3$	$Cmcm$	0.06–0.71	1.51–2.13	1.08–1.50	0.008	0.002	no	4
mp-760655–mp-757905	$Li_3(Ti_xBi_{1-x})(PO_4)_2$	$C2/m$	0.44–0.60	1.51–2.08	1.16–1.48	0.066	0.072	yes	6
mp-18903–mp-18848	$Sr_2(Mg_xCd_{1-x})WO_6$	$Fm\bar{3}m$	0.16–0.54	3.39–3.50	0.95–1.50	0.082	0.009	no	0
mp-18848–mp-19400	$Sr_2(Ni_xMg_{1-x})WO_6$	$Fm\bar{3}m$	0.5–0.53	1.52–1.65	1.45–1.50	0.009	0.010	no	0

†Whether a composition within x range lies on the half-space hull. ‡Number of decomposition products from half-space hull; excludes endpoint compounds from count.

incommensurate alloy pairs, and each of the cases mentioned here could be incorporated into future iterations of this alloys database.

We note that the underlying, input database from which our alloy database is derived can contain biases. These biases, e.g. concerning structural as well as chemical coverage, can propagate into the alloy database, which should be acknowledged when interpretting results. As the underlying database expands, this infrastructure has been established to automatically "build" new versions of the alloy database as new data becomes available. Importantly, as better methods for calculating more accurate lattice parameters or band gaps become accessible for high-throughput computation, the alloy database incorporates this improved data.

Once a set of potential alloys are suggested from this database, more reliable methods to assess alloy solubility can be used to either rule out or confirm a potential alloy. For example, automated cluster expansions[38] or the generalized quasi-chemical approximation (GQCA) method.[39] Our work is intended to serve as a starting point from which to determine systems to consider for such in-depth analyses. The half-space hull diagrams provide a guide to select alloys within a given chemical space which may be stable and synthesizable. For example, the following calculations of increasing computational cost could be explored based on outputs from the alloys database:

- For compounds at endpoint A (or B) in which a commensurate compound at endpoint B (or A) is not present on MP, there is insufficient information in the database to calculate an alloy pair (for example, the black circular markers in Figure 3 without tielines) such as $Zn_xCu_{1-x}S$). Here, the missing compound(s) can be calculated and added to the database. This is still important even if such a compound is unstable or not experimentally realizable at the endpoint, since there may be a region within alloy space where synthesizability becomes possible.

- For alloy pairs in which member compounds are not yet known to exist, members can be calculated (e.g. at x=0.5) for a few different orderings to assess realizability, or give an indication of expected bowing and other parameters.

- Many real alloy materials are *disordered*, rather than ordered. For members within in alloy pair, special quasi-random structure (SQS) calculations can approximate structures of fully random alloy polymorphs to provide a counterpoint to the small-cell ordered structures more typical in a database such as the Materials Project.[40]

- To account for configurational entropy and thermodynamics of specific alloy members, the generalized quasi-chemical approximation (GQCA) can be used to estimate free energy,[39] and subsequently higher order methods such as cluster expansions can be applied to further investigate specific systems for which high quality phase diagrams are required.[38]

For immediate use, our alloy database has been incorporated into the Materials Project as an app in the new website release and API, in the hope that this will serve as a guide for researchers performing screenings of tunable materials. The alloys database will be updated alongside the Materials Project database. A flowchart of the alloys database pipeline and incorporation onto the Materials Project is shown in **Figure 6**.

CONCLUSION

In this paper we have presented a new framework to analyze alloys in the context of materials databases, implemented it into the open source

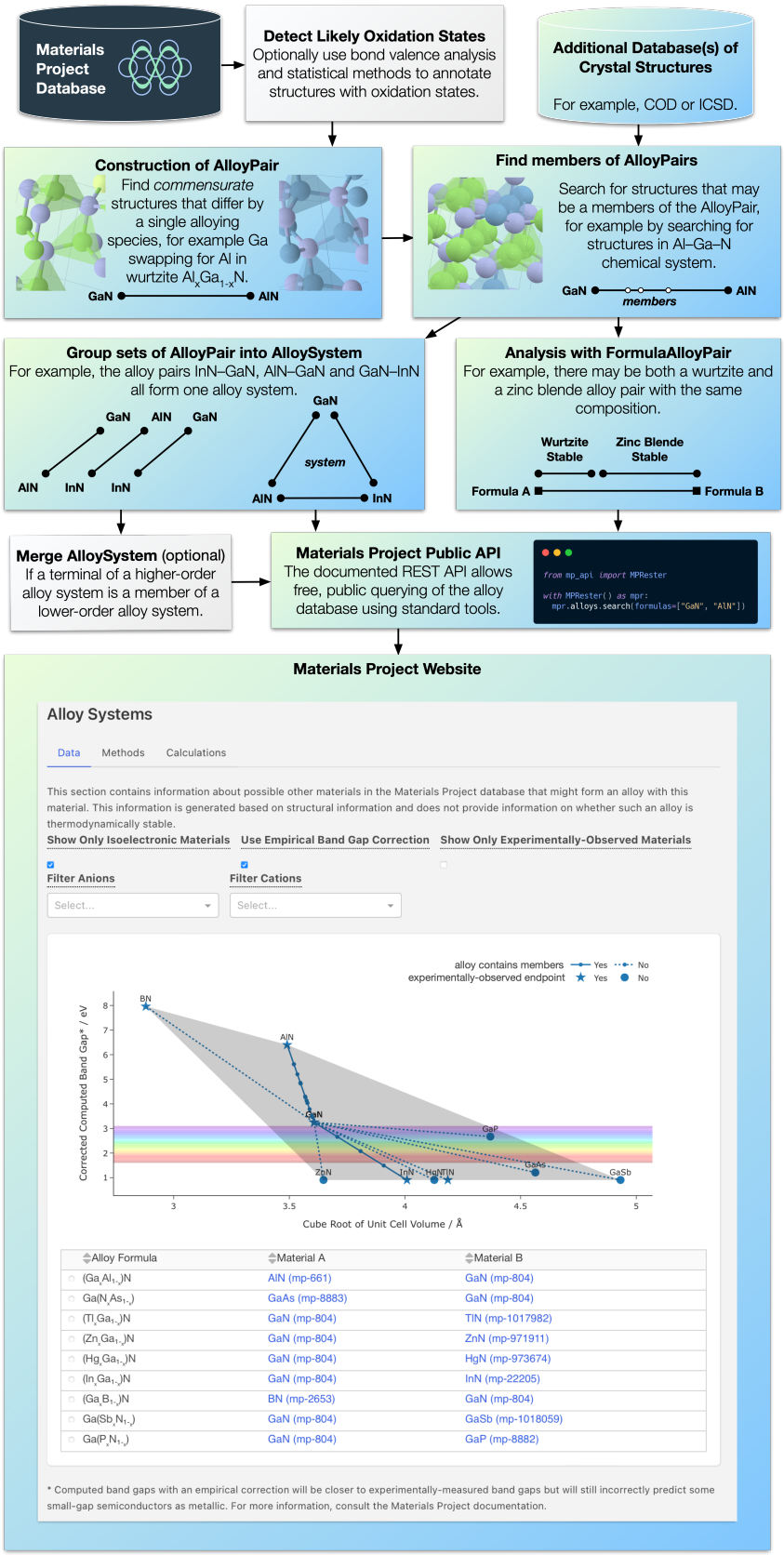

Figure 6: A flowchart showing the data processing pipeline outlined in the methodology, starting from a generic crystal structure database such as the Materials Project, and ending with a publicly accessible API and website to explore the data. Wurtzite GaN is shown here as an example, as an alloy pair with InN and as an alloys system plotted on the MP website.

`pymatgen-analysis-alloys` package, and created an open-source alloys database that has been incorporated into the Materials Project website. We have presented a few case studies here of how this database can be utilized in the context of materials research and design.

Importantly, all the analysis presented here has been performed without any new calculations, which showcases some of the data analysis opportunities from mining existing databases. A decade into the Materials Genome Initiative, the materials discovery community has produced large quantities of data in multiple databases, but data *production* is just the start; it is essential that data is curated, structured, and connected in a way to yield the maximum value to the community.

In particular, one of the key challenges is how to link and apply this data to successfully use computational predictions to inform experimental results, especially as experimental databases grow.[41, 42] In particular, experimental progress in semiconductors typically starts from a well-studied, well-characterized material and modifies its properties iteratively with the addition of dopants or alloying elements during growth. The framework of this paper addresses this aspect of materials design by creating a database of candidate, tunable materials by a data-focused approach which can use existing materials databases to suggest alloys between pairs of already-known materials. Thus, a new materials screening procedure is now possible that can emphasize experimentally-accessible materials and suggest screening outputs that would have been previously wholly overlooked.

METHODOLOGY

We have created an open-source code, `pymatgen-analysis-alloys`, that allows the construction of an alloy database when provided with an input database containing crystal structures. As a demonstration, we apply this code to the Materials Project database. The left side of Figure 6 depicts the data processing pipeline of the alloys database, as described here. This code is also used for the automatic generation of the plots shown in this manuscript, with only light additional editing performed for presentation.

The method outlined here does not require any prior knowledge of which materials might form alloys. While partial occupancies in e.g. a Crystallographic Information File (.cif) indicates the possibility of alloying, this criteria only captures known systems, and hence does not fully explore the possible alloy space. The challenge when constructing the database is in the data processing pipeline, and addressing combinatorial problems when large databases of hundreds of thousands of entries are used.

The method is as follows: for each crystal structure in the input database, designated as a potential "endpoint," we find all other compounds that share its anonymous formula (e.g. "ABC_2"), and perform a pairwise comparison between all materials to detect commensurate structures using the `StructureMatcher` in `pymatgen`[43]. A pre-filter is performed that checks for detected space group, calculated with `spglib`[44], using both tight and loose tolerances. This pre-filter is imposed with the logic that it is a necessary but not sufficient condition that two commensurate crystal structures will have the same space group. After a pair of crystal structures are identified as an endpoint, information is extracted such as the alloying species, including oxidation state, and whether the alloy is isoelectronic, and stored as an instance of an `AlloyPair` class. This definition of "alloy" does not consider alloys formed through interstitial alloying additions or other types of alloys.

All `AlloyPair` entries contain structural properties, such as space group and primitive cell volume, but can be supplemented with additional properties. For this demonstration, supplemental properties are taken from the Materials Project and include E_{hull}, E_{G} from the Purdue-Berke-Ernzerhof (PBE) functional from the Materials Project, and electron and hole effective masses from Ricci et al.[35] (note that these are only computed for a subset of the MP database), but in principle this can be expanded to include any material property. Methods are provided to interpolate these properties using Vegard's law (assuming no bowing) for a given alloy content to allow for easier plotting and searching (see SI).

Once a set of `AlloyPair` entries are constructed, they are grouped by chemical system and iterated over to search for potential members, defined by the `AlloyMember` class, using a similar approach. This allows a database query to reveal which alloys already have existing data available, and thus may be more experimentally-accessible alloys when performing a screening.

A set of alloy pairs can be grouped together as an "alloy system," defined by the class `AlloySystem`, using a network graph method whereby each edge in the graph is an alloy pair and connected subgraphs form the respective alloy systems. The code allows for alloy systems to be merged when a member of one system might be the endpoint of another system, for example an alloy system with ternary endpoints where one endpoint is itself a member of a binary alloy pair.

Another useful grouping is the set of alloy pairs which all have the same set of endpoint formulae: these can be grouped together as a "formula alloy pair," defined by the class `FormulaAlloyPair`. If formation enthalpies are known for the endpoints, this class is able to define regions where a given polymorph is stable according to a simple linear interpolation, and define alloy segments (class `AlloySegment`) which encode the critical alloy contents at which a phase transition may occur between two polymorphs. Furthermore, if any alloy members are known, including their formation enthalpies, this data will inform how accurate the simple linear interpolation may be. Examples of these can be seen in Figure 4.

For the example database generated in this work, we exclude compounds including H, He, noble gases, and heavy elements with atomic numbers greater than 83 (Bi),

although all these entries are present in the underlying database, but we do not perform any further filtering based on chemistry and leave this as a capability for the user querying the database to decide exactly what chemical systems, maximum electronegativity differences, etc. are allowable for their specific design case.

The API to access and search the database is defined in the open-source `emmet` code. The user interface on the Materials Project website is constructed using the open-source Crystal Toolkit web framework. All open-source code described in this work is open to review and suggested edits by other researchers, and any contributions are welcomed by the authors.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (Materials Project program KC23MP). R.W.R. acknowledges financial support from the U.C. Berkeley Chancellor's Fellowship and the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. DGE1106400 and DGE175814.

AUTHOR CONTRIBUTIONS

We highlight the author contributions to this study using the CRediT taxonomy.

R.W.R: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.

M.K.H.: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.

K.A.P.: Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing.

[1] T. Mukai, M. Yamada, and S. Nakamura, "Characteristics of ingan-based uv/blue/green/amber/red light-emitting diodes," *Japanese Journal of Applied Physics*, vol. 38, no. 7R, p. 3976, 1999.

[2] M. A. Kinch, "Fundamental physics of infrared detector materials," *Journal of Electronic Materials*, vol. 29, no. 6, pp. 809–817, 2000.

[3] S. R. Anton and H. A. Sodano, "A review of power harvesting using piezoelectric materials (2003–2006)," *Smart Materials and Structures*, vol. 16, no. 3, p. R1, 2007.

[4] N. G. Dhere, "Present status and future prospects of CIGSS thin film solar cells," vol. 90, no. 15, pp. 2181–2190.

[5] A. E. Rubin, "Mineralogy of meteorite groups," *Meteoritics & Planetary Science*, vol. 32, no. 2, pp. 231–247, 1997.

[6] H. Zhu, G. Hautier, U. Aydemir, Z. M. Gibbs, G. Li, S. Bajaj, J.-H. Pöhls, D. Broberg, W. Chen, A. Jain, *et al.*, "Computational and experimental investigation of $TmAgTe_2$ and XYZ_2 compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening," *Journal of Materials Chemistry C*, vol. 3, no. 40, pp. 10554–10565, 2015.

[7] H. Chen, Q. Hao, O. Zivkovic, G. Hautier, L.-S. Du, Y. Tang, Y.-Y. Hu, X. Ma, C. P. Grey, and G. Ceder, "Sidorenkite ($Na_3MnPO_4CO_3$): a new intercalation cathode material for na-ion batteries," *Chemistry of Materials*, vol. 25, no. 14, pp. 2777–2786, 2013.

[8] G. Hautier, A. Miglio, G. Ceder, G.-M. Rignanese, and X. Gonze, "Identification and design principles of low hole effective mass p-type transparent conducting oxides," *Nature communications*, vol. 4, no. 1, pp. 1–7, 2013.

[9] R. Woods-Robinson, D. Broberg, A. Faghaninia, A. Jain, S. S. Dwaraknath, and K. A. Persson, "Assessing high-throughput descriptors for prediction of transparent conductors," vol. 30, no. 22, pp. 8375–8389.

[10] R. Li, L. Xie, W. Y. Wang, P. K. Liaw, and Y. Zhang, "High-throughput calculations for high-entropy alloys: A brief review," *Frontiers in Materials*, vol. 7, Sept. 2020.

[11] Y. Lederer, C. Toher, K. S. Vecchio, and S. Curtarolo, "The search for high entropy alloys: A high-throughput ab-initio approach," *Acta Materialia*, vol. 159, pp. 364–383, Oct. 2018.

[12] K. C. Pitike, S. KC, M. Eisenbach, C. A. Bridges, and V. R. Cooper, "Predicting the phase stability of multicomponent high-entropy compounds," *Chemistry of Materials*, vol. 32, pp. 7507–7515, July 2020.

[13] S. Jiang and K. Yang, "Review of high-throughput computational design of heusler alloys," *Journal of Alloys and Compounds*, vol. 867, p. 158854, 2021.

[14] S. Sanvito, C. Oses, J. Xue, A. Tiwari, M. Zic, T. Archer, P. Tozman, M. Venkatesan, M. Coey, and S. Curtarolo, "Accelerated discovery of new magnets in the heusler alloy family," *Science Advances*, vol. 3, Apr. 2017.

[15] C. A. C. Garcia, J. D. Bocarsly, and R. Seshadri, "Computational screening of magnetocaloric alloys," *Physical Review Materials*, vol. 4, Feb. 2020.

[16] S. Bhattacharya and G. K. H. Madsen, "High-throughput exploration of alloying as design strategy for thermoelectrics," *Physical Review B*, vol. 92, Aug. 2015.

[17] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations," *Computational Materials Science*, vol. 58, pp. 227–235, June 2012.

[18] S. Kirklin, J. E. Saal, V. I. Hegde, and C. Wolverton, "High-throughput computational search for strengthening precipitates in alloys," *Acta Materialia*, vol. 102, pp. 125–135, Jan. 2016.

[19] T. Bligaard, G. H. Jóhannesson, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov, "Pareto-optimal alloys," *Applied Physics Letters*, vol. 83, pp. 4527–4529, Dec. 2003.

[20] K. Yang, C. Oses, and S. Curtarolo, "Modeling off-stoichiometry materials with a high-throughput ab-initio approach," *Chemistry of Materials*, vol. 28, pp. 6484–6492, Sept. 2016.

[21] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, "Accelerating high-throughput searches for new alloys with active learning of interatomic potentials," *Computational Materials Science*, vol. 156, pp. 148–156, Jan. 2019.

[22] A. van de Walle and M. Asta, "High-throughput calcula-

tions in the context of alloy design," *MRS Bulletin*, vol. 44, pp. 252–256, Apr. 2019.
[23] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nature Materials*, vol. 12, pp. 191–201, Feb. 2013.
[24] U. Mizutani, "Hume-rothery rules for structurally complex alloy phases," *Mrs Bulletin*, vol. 37, no. 2, pp. 169–169, 2012.
[25] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL materials*, vol. 1, no. 1, p. 011002, 2013.
[26] A. M. Holder, S. Siol, P. F. Ndione, H. Peng, A. M. Deml, B. E. Matthews, L. T. Schelhas, M. F. Toney, R. G. Gordon, W. Tumas, *et al.*, "Novel phase diagram behavior and materials design in heterostructural semiconductor alloys," *Science advances*, vol. 3, no. 6, p. e1700270, 2017.
[27] G. Bergerhoff and I. D. Brown, "Crystallographic databases/Allen FH et al.(hrsg.). Chester, international union of crystallography,"
[28] J. Singh, *Electronic and optoelectronic properties of semiconductor structures*. Cambridge University Press, 2007.
[29] J. Piprek, *Semiconductor optoelectronic devices: introduction to physics and simulation*. Elsevier, 2013.
[30] R. Woods-Robinson, Y. Han, H. Zhang, T. Ablekim, I. Khan, K. A. Persson, and A. Zakutayev, "Wide band gap chalcogenide semiconductors," vol. 120, no. 9, pp. 4007–4055.
[31] Á. Morales-García, R. Valero, and F. Illas, "An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations," *The Journal of Physical Chemistry C*, vol. 121, no. 34, pp. 18862–18866, 2017.
[32] A. Banerjee and K. Chattopadhyay, "Recent developments in the emerging field of crystalline p-type transparent conducting oxide thin films," vol. 50, no. 1-3, pp. 52–105.
[33] N. Sarmadian, R. Saniz, B. Partoens, and D. Lamoen, "Easily doped p-type, low hole effective mass, transparent oxides," vol. 6, pp. 1–9.
[34] J. B. Varley, A. Samanta, and V. Lordi, "Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels," vol. 8, no. 20, pp. 5059–5063.
[35] F. Ricci, W. Chen, U. Aydemir, G. J. Snyder, G.-M. Rignanese, A. Jain, and G. Hautier, "An ab initio electronic transport database for inorganic materials," *Scientific data*, vol. 4, p. 170085, 2017.
[36] V. Stevanović, R. Trottier, C. Musgrave, F. Therrien, A. Holder, and P. Graf, "Predicting kinetics of polymorphic transformations from structure mapping and coordination analysis," *Physical Review Materials*, vol. 2, no. 3, p. 033802, 2018.
[37] F. Therrien, P. Graf, and V. Stevanović, "Matching crystal structures atom-to-atom," *The Journal of chemical physics*, vol. 152, no. 7, p. 074106, 2020.
[38] D. B. Laks, L. Ferreira, S. Froyen, and A. Zunger, "Efficient cluster expansion for substitutional systems," *Physical Review B*, vol. 46, no. 19, p. 12587, 1992.
[39] A.-B. Chen and A. Sher, *Semiconductor alloys: physics and materials engineering*. Springer Science & Business Media, 1995.
[40] A. Zunger, S.-H. Wei, L. Ferreira, and J. E. Bernard, "Special quasirandom structures," vol. 65, no. 3, p. 353.
[41] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, and C. Phillips, "An open experimental database for exploring inorganic materials," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.
[42] K. R. Talley, R. White, N. Wunder, M. Eash, M. Schwarting, D. Evenson, J. D. Perkins, W. Tumas, K. Munch, C. Phillips, *et al.*, "Research data infrastructure for high-throughput experimental materials science," *Patterns*, vol. 2, no. 12, p. 100373, 2021.
[43] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," *Computational Materials Science*, vol. 68, pp. 314–319, 2013.
[44] A. Togo and I. Tanaka, "`spglib`: a software library for crystal symmetry search," *arXiv preprint arXiv:1808.01590*, 2018.
[45] H. Yang, T. Song, X. Liang, and G. Zhao, "First-principle study of the electronic band structure and the effective mass of the ternary alloy gaxin1-xp," in *Journal of Physics: Conference Series*, vol. 574, p. 012048, IOP Publishing, 2015.

Supplementary Information

Open-source code

All code used in the preparation of this manuscript is open source. Where future developments require changes to the methods or algorithms described in this manuscript, these open source codes will contain the ground truth for how the alloy database is constructed.

The codes developed were:

- `pymatgen-analysis-alloys` An add-on package for the `pymatgen` code that contains the `AlloyPair`, `AlloyMember`, `AlloySystem` and `FormulaAlloyPair` classes and related logic.
- `emmet` This is an existing package containing information on how to build the databases used by the Materials Project. Code was added to `emmet-core` to define the database document schema and `emmet-builders` to define the scripts to construct the database in a scalable manner. Code was added to `emmet-api` to allow researchers to access the alloy database constructed in this work through the Materials Project.

At the time of writing, `pymatgen-analysis-alloys` is installable using the Python Package Index via `pip install pymatgen-analysis-alloys` and importable via `import pymatgen.analysis.alloys`. The main classes are located in `pymatgen.analysis.alloys.core` and are documented and unit tested. Readers are encouraged to refer to the code for any updates to this methodology subsequent to publication.

Unique Identifiers

This work uses a document-based database, namely MongoDB, which does not have an explicit schema. The database fields present will be derived based on the available attributes in the `AlloyPair` and other objects. The canonical reference for these attributes is the code itself.

Nevertheless, the use of a unique, primary key is essential for database management.

For `AlloyPair` this is an underscore-delimited string containing the unique identifiers of the endpoints from whatever input database is used. This implicitly assumes that an underscore is not used in the input databases's unique identifiers. For example, an `AlloyPair` consisting of materials mp-804 (GaN) and mp-661 (AlN) would have the unique identifier "mp-661_mp-804". The `AlloyPair` construction orders the endpoints deterministically, such that AlN will always be endpoint "A" and GaN will always be endpoint "B" regardless of the order of endpoints provided during construction.

For `AlloySystem`, the unique identifier is based on the first six digits of the MD5 hash of a sorted, underscore-delimited list of all unique identifiers of individual materials in that alloy system. This ensures that the identifier will change as additional members are added to the alloy system.

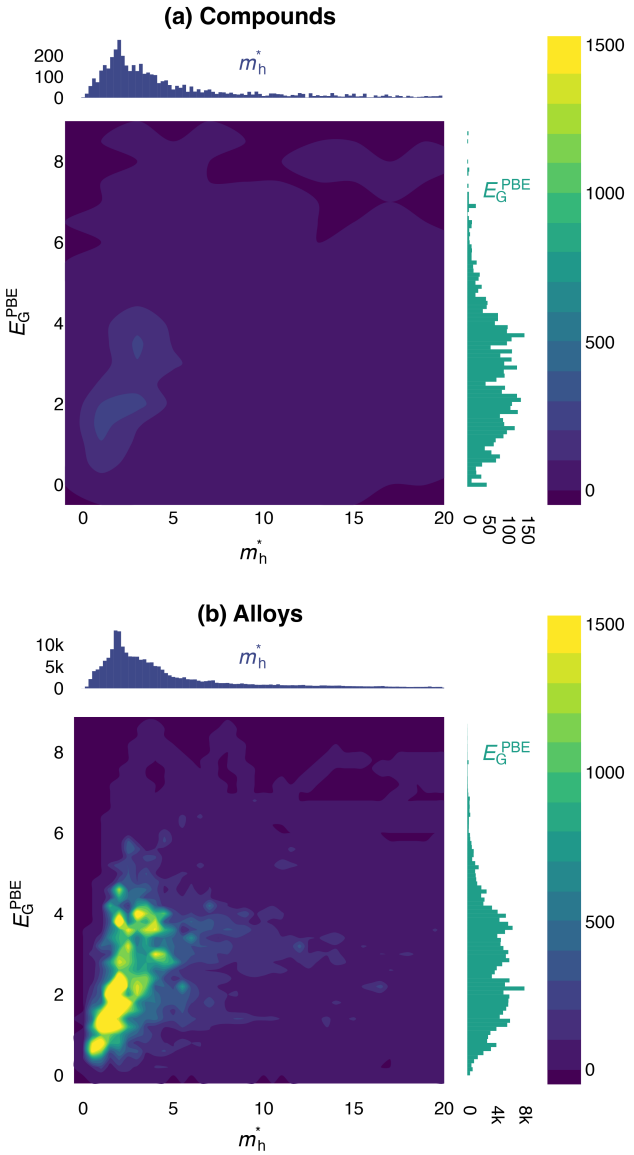

Figure S1: 2D density plots showing (a) the distribution of E_G and m_h^* when considering only stoichiometric alloy compounds ("endpoints") and (b) an approximate distribution of E_G and m_h^* including the intermediate alloy compositions, illustrating the different between a discrete and practically continuous distribution of properties.

Database Building

Constructing the entire alloy database is CPU-bound and takes approximately one day on a 2.3 GHz 8-core Intel CPU. For Materials Project production purposes, this database build is typically parallelized across multiple nodes and "pleasingly parallel", since it can be parallelized across anonymous formula (for alloy pair and alloy system construction) and across chemical system (for alloy member construction) such that the total build time is greatly reduced.

Merging of AlloySystem

Consider there is an alloy system containing an endpoint with anonymous formula ABC. However, this endpoint ABC is also found to be a member of another alloy system (say, a system made up of the two endpoints, AB and AC). In this case, we can conclude the first alloy system can be subsumed into the second alloy system. This process of "merging" alloy systems is important to remove spurious alloy systems, but also presents a subtle problem since whether a material should be considered an alloy or, simply, a new stoichiometric compound is open to interpretation. For example, chalcopyrite is typically considered a compound in its own right, but under this lens would be seen as an alloy of two zincblende endpoints. Therefore, alloy system merging has not been performed on the database in this work, but has been fully implemented in the code and can be done manually on an as-needed basis.

Vegard's Law Approximations

In the manuscript, we assume Vegard's law applies with no bowing to construct Figure S1, Figure 5, and for properties a, E_G, and inverse effective mass (i.e. $\frac{1}{m_\mathrm{e}^*}$ and $\frac{1}{m_\mathrm{h}^*}$). This is a crude approximation for the purposes of providing a window for a given alloy in which properties might lie, *not* as a way to accurately estimate properties. The literature commonly applies Vegard's law for alloys to estimate a and E_G. In comparison, there is less consensus across the literature about whether Vegard's law is appropriate for m^* and whether bowing is pronounced; this is likely dependent on the specific characteristics of the electronic band structure for the alloy endpoints. According to Piprek[29] and Singh[28], Vegard's law is appropriate for inverse effective mass with the latter providing a derivation. According to Piprek "bowing is not pronounced for the effective mass of most alloys," as compared to stronger bowing for band gap. We note that "most alloys" is likely referring to III-V materials, since these are the dominant class of alloys studied — and for III-Vs, Vegard's law is used in the literature to estimate effective mass e.g. for (AlGaIn)N alloys.[45]

Alloys increase parameter space

To graphically illustrate how including alloys increases parameter space, **Figure S1** depicts a 2D contour plot of two representative material properties — m_h^* versus PBE E_G (see Methodology) — for (a) compounds in the MP database, i.e. endpoints only, in comparison to (b) candidate alloy materials with steps of $\delta x = 0.01$ in an alloy pair $\mathrm{A}_x\mathrm{B}_{1\text{-}x}$ and assuming Vegard's law with no bowing for E_G and $1/m_\mathrm{h}^*$.[28] The histograms above and to the right of each diagram depict the distribution for each individual parameter. Note that this is simply an illustration, to show the expanded property space accessible when considering alloys, and is not a quantitative comparison since the choice of δ is arbitrary.